Response surface methods for selecting spectrogram hyperparameters with application to acoustic classification of environmental-noise signatures

Science of Test Workshop
Waterford at Springfield, April 4th, 2017

Ed Nykaza (ERDC-CERL)
Pete Parker (NASA-Langley)
Matt Blevins (ERDC-CERL)
Anton Netchaev (ERDC-ITL)
Environmental Noise is a Problem
Sleep Disturbance – Annoyance – Health Effects
Environmental Noise is Ubiquitous
Continuous – Impulsive – Intermittent Noise Sources
Manage Environmental Noise with Continuously Recording Noise Monitors

JUST TWO PROBLEMS
JUST TWO PROBLEMS … (OK, MAYBE THREE)

1. WAY TOO MUCH DATA!
   – Each monitoring station may record millions of events over the course of a year

2. NEED TO KNOW WHAT WAS RECORDED!
   – Too much data for humans to listen to

SOLUTION: BUILD AN ENVIRONMENTAL NOISE CLASSIFIER THAT CAN CLASSIFY ALL ENVIRONMENTAL NOISE SOURCES
   – We consider using spectrograms as they have worked well in many music and speech classification problems
   – BUT, spectrograms have many input parameters that can take many values (AKA hyperparameters)

3. HOW DO WE CHOOSE THE BEST SPECTROGRAM HYPERPARAMETERS THAT WILL GIVE THE MOST ACCURATE ENVIRONMENTAL NOISE CLASSIFIER?
The General Problem

- How do we choose “best” hyperparameters to give us the “best” results?
- What if we have many inputs?
- What if each input has many possible values?
- What if it is not computationally feasible to try each combination of hyperparameters?

- **In Summary:** What hyperparameters values should we choose to achieve adequate accuracy with minimal computational cost?
Environmental Noise Classification Problem

Input Signal(s) → Feature Extraction Method(s) → Classification Model(s) → Predict which Sounds were Recorded
A study of the relationship between spectrogram hyperparameters and classification accuracy

- Human Listener
- Labeled Dataset
- Input Signals
- Spectrogram (as features)
- Support Vector Machine Classifier
- Classification Accuracy

- 6 input parameters
- Varying number of possible values per each parameter
- 284+ million unique combinations
**NYU Urban Sound 8k Dataset**

- 7 classes x 450 sound recordings each = 3150 total
  - Air conditioner (y1)
  - Children playing (y2)
  - Dog barking (y3)
  - Drilling (y4)
  - Engine idling (y5)
  - Jackhammer (y6)
  - Street music (y7)
Spectrogram As Features

- **Spectrogram**
  - Squared magnitude of short-time Fourier transform (STFT)
  - Acoustic intensity vs. time and frequency
• **Hyperparameters considered:**
  - nFFT – number of points in FFT
    - nFFT $\epsilon [2^N]$, where $N = 1 - 8$
  - nTimeBins – number of time windows
    - nTimeBins $\epsilon [2^N]$, where $N = 1 - 8$
  - %Overlap – percentage of time window overlap
    - Overlap $\epsilon [0, 99]$
  - A total of 6400 unique combinations ($8 \times 8 \times 100$)

• **Hyperparameters NOT considered in order to make problem more tractable:**
  - Window-types (used Kaiser window in this study)
  - Signal duration (truncated all signals to 2 seconds)
  - Sample rate (down-sampled to 4096 Hz)
Classification & Accuracy with SVM

- Multi-class support vector machine (SVM with radial basis function kernel and automatic scale)

- 5-fold cross-validation (train on 4 folds, test on 1)

- Overall Accuracy (% correctly classified, all classes)

- Class Accuracy (% correctly classified per class)
Study Summary & Analysis Plan

NYU Urban Sound Dataset

Spectrogram (as features)

Support Vector Machine Classifier

Hyperparameter Sampling Methods

Spectrogram Hyperparameters (6400 unique combinations)

• Response Surface Method (RSM)
  • Compare to:
    • Exhaustive Sample (6400 points)
    • Other Sampling Methods
      • Genetic Algorithms
      • Particle Swarm
      • Differential Evolution

• Overall Accuracy
• Class Accuracy
  • y4 (drilling)
  • y6 (jackhammer)
Cast Problem in Response Surface Method (RSM) Framework

- **Factors**
  - $x_1$: nFFT
  - $x_2$: nTimeBins
  - $x_3$: %Overlap

- **Parameters**
  - low (-1)
  - high (+1)

<table>
<thead>
<tr>
<th>Factors</th>
<th>Parameters</th>
<th>low (-1)</th>
<th>high (+1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>nFFT</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>$x_2$</td>
<td>nTimeBins</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>$x_3$</td>
<td>%Overlap</td>
<td>0</td>
<td>99</td>
</tr>
</tbody>
</table>

- **Responses (y’s)**
  - Classification accuracy (% correct) of 7 Sound Classes ($y_1$-$y_7$)
  - Overall accuracy, $y_8$ average performance across all classes
  - Computational Cost ($z_1$), measured by solution time in (sec)
• **Classical geometric design**
  – 25 total design points
  – 14 face-centered central composite design (O)
  – 8 inscribed, interior factorial points (O)
  – 3 Center points to assess variability in 5-fold cross validation
• 5 unique levels of each factor
• Design points are approximate geometric locations for x1 and x2, since they are restricted to be integers
RSM Assumed Model Form

• **A typical RSM approach assumes at Taylor Series model**
  – Empirical, meta-modeling approach
  – Flexible, graduated function
  – Linear in the coefficients, allowing for linear regression
  – Proven useful/sufficient in research and product optimization

• **Second-order Model**

\[
y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j
\]

  – \( y \) is the response (\( z \) for cost)
  – \( x \)'s are the factors, often coded/normalized (\(+/- \ 1\))
  – \( k \) is the number of factors
  – **betas** are the regression coefficients to be estimated
  – additive error term, not shown
## RSM Modeling

### Overall Accuracy (y8)

### Sorted Parameter Estimates

| Term                              | Estimate | Prob>|t| |
|-----------------------------------|----------|-----|
| nFFTs (x1)                        | 0.109    | <.0001* |
| nTimeBins (x2)                    | 0.010    | 0.0964 |
| nFFTs * % Overlap (x1*x3)         | 0.014    | 0.1091 |
| % Overlap (x3)                    | 0.010    | 0.1965 |
| nTimeBins * nTimeBins (x2^2)      | -0.015   | 0.2350 |
| nFFTs * nFFTs (x1^2)              | -0.011   | 0.3485 |
| % Overlap * % Overlap (x3^2)      | -0.017   | 0.3495 |
| nTimeBins * % Overlap (x2*x3)     | 0.008    | 0.3521 |
| nFFTs * nTimeBins (x1*x2)         | -0.001   | 0.9196 |
Cube Plot
Overall Accuracy y8

19
of 38
How does RSM Overall Accuracy compare to exhaustive sample?

\[ n_{FFT} \leq 4 \]

\[ n_{FFT} \leq 3 \]

\[ n_{FFT} \leq 6 \]

\[ n_{FFT} \leq 7 \]
Computational Cost ($z_1$) Model

\[ \ln(z_1) = \text{Comp Time} \]

- $x_1$ (nFFTs)
- $x_2$ (nTimeBins)
- $x_3$ (%) Overlap

Graph showing the relationship between computational cost and the factors $x_1$, $x_2$, and $x_3$. The graph includes data points and lines indicating the increase in computational cost as these factors increase.
How does RSM Computational Time compare to exhaustive sample?
RSM Summary
Overall Accuracy (y8)

• Conclusion: X1 (nFFT) is the most significant effect for overall performance (agrees with exhaustive sample)

• A linear expression can predict the overall accuracy with a stddev of ~2.5%  
  \[ y_8 = 0.3728 + 0.1104 \times \frac{(x_1 - 4.5)}{3.5} \]
  - Increase x1(nFFT) to increase overall classification accuracy
  - x2 (nTimeBins) can be set low to minimize computational cost with little effect on overall classification accuracy

• Repeating analysis for individual sound classes, yields different results and interesting insights…
RSM Significant Terms by Sound Class

- x3 (% Overlap)
- $x2 \times x3$ (nTimeBins x %Overlap)
- $x2 \times x2$ (nTimeBins^2)
- x2 (nTimeBins)
- $x1 \times x2$ (nFFT x nTimeBins)
- $x1 \times x1$ (nFFT^2)
- x1 (nFFT)

Terminology:
- y1 (Air Cond)
- y2 (Children Playing)
- y3 (Dog Barking)
- y4 (Drilling)
- y5 (Engine Idling)
- y6 (Jack Hammer)
- y7 (Street Music)
- y8 (Average)_Avg

Legend:
- abs(estimate)

Values:
- 0.25
- 0.20
- 0.15
- 0.10
- 0.05 Estimate
- 0.00
- -0.05
- -0.10
- -0.15
Inferences from Significant Term Plot

• Utility of spectrogram hyperparameters vary by sound class

• Some parameters are inversely proportional
  – Lower value results in better classification
  – And typically less computational cost

• x3 (% Overlap) has overall weak effect
  – Exceptions are y5 (engine idling) and y6 (jackhammer)
  – Implies that x3 can be set to minimize computational costs
  – Implies that x3 does not need to vary every percent
Inferences from Significant Term Plot

• **y6 (Jack Hammer)** indicates that x1 (nFFT) is not an important factor, rather x2 (nTimeBins) is the dominant effect & inversely proportional

• **y4 (Drilling)** indicates that there is no strong relationship between the spectrogram hyperparameters and the classification accuracy…**but** let’s take a closer look…
Let’s Take a Closer Look at y4 (Drilling) Contour using all 25 RSM Points
Captured in 4th Order RSM Model

| Term                                                                 | Prob>|t|  |
|----------------------------------------------------------------------|------|---|
| x₁ (nFFTs) * x₂ (nTimeBins) * x₂ (nTimeBins)                        | 0.0014* |   |
| x₁ (nFFTs) * x₂ (nTimeBins)                                         | 0.0022* |   |
| x₂ (nTimeBins) (1,8)                                                | 0.0022* |   |
| x₁ (nFFTs) (1,8)                                                    | 0.0035* |   |
| x₁ (nFFTs) * x₃ (% Overlap) * x₃ (% Overlap)                         | 0.0040* |   |
| x₃ (% Overlap) * x₃ (% Overlap)                                      | 0.0051* |   |
| x₁ (nFFTs) * x₁ (nFFTs)                                             | 0.0139* |   |
| x₁ (nFFTs) * x₁ (nFFTs) * x₂ (nTimeBins) * x₂ (nTimeBins)           | 0.0148* |   |
| x₁ (nFFTs) * x₁ (nFFTs) * x₂ (nTimeBins)                            | 0.0497* |   |
| x₃ (% Overlap) (0,99)                                                | 0.0857 |   |
| x₂ (nTimeBins) * x₂ (nTimeBins)                                     | 0.1496 |   |

* Terms above 2nd Order
Contour Plot Comparison
25 RSM vs. 4th Order RSM Model

All 25 RSM Points

Predicted
4th Order RSM Model

Note: showing average accuracy, averaged over all %Overlaps
Contour Plot Comparison
25 RSM vs. Exhaustive Sample

Note: showing average accuracy, averaged over all %Overlaps
Other Sampling (Optimization) Methods

- **Genetic Algorithm**
  - Local search algorithm that mimics evolution
  - Developed by J. Holland, K. DeJong, D. Goldberg in 1970’s
  - Typically applied for discrete optimization

- **Differential Evolution**
  - Simple and efficient adaptive scheme for global optimization over continuous spaces
  - Proposed by Price and Storn in 1995

- **Particle Swarm**
  - Inspired by nature’s social behavior and dynamic movements with communications of animals such as insects, birds, and fish
  - Developed by Dr. Ebehart and Dr. Kennedy in 1995
Particle Swarm Method
Best Hyperparameters

• Recall, the primary question is “What hyperparameters values should we set to achieve adequate accuracy with minimal computational cost?”

– Varies by sound class!

– However, RSM found them with only 25 samples! Recall that we knew the answer with the exhaustive sample.
# Best Hyperparameters with RSM

- **y8 (Overall Accuracy)**

<table>
<thead>
<tr>
<th>Sampling Method</th>
<th>nFFT (x1)</th>
<th>nTimeBins (x2)</th>
<th>%Overlap (x3)</th>
<th>Accuracy</th>
<th>Runtime (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSM</td>
<td>high as poss. ($2^8$)</td>
<td>low for comp. ($2^1$)</td>
<td>any value (50)</td>
<td>47.2%</td>
<td>11</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>256 ($2^8$)</td>
<td>8 ($2^3$)</td>
<td>16%</td>
<td>51.2%</td>
<td>23</td>
</tr>
</tbody>
</table>

- **y6 (Jackhammer)**

<table>
<thead>
<tr>
<th>Sampling Method</th>
<th>nFFT (x1)</th>
<th>nTimeBins (x2)</th>
<th>%Overlap (x3)</th>
<th>Accuracy</th>
<th>Runtime (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSM</td>
<td>low for comp. ($2^1$)</td>
<td>2 ($2^1$)</td>
<td>90%</td>
<td>48.0%</td>
<td>8</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>2 ($2^1$)</td>
<td>32 ($2^5$)</td>
<td>90%</td>
<td>64.0%</td>
<td>13</td>
</tr>
</tbody>
</table>

- **y4 (Drilling)**

<table>
<thead>
<tr>
<th>Sampling Method</th>
<th>nFFT (x1)</th>
<th>nTimeBins (x2)</th>
<th>%Overlap (x3)</th>
<th>Accuracy</th>
<th>Runtime (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSM</td>
<td>32 ($2^5$)</td>
<td>2 ($2^1$)</td>
<td>any value (50)</td>
<td>61.3%</td>
<td>8</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>32 ($2^5$)</td>
<td>16 ($2^4$)</td>
<td>90%</td>
<td>65.6%</td>
<td>13</td>
</tr>
</tbody>
</table>
Summary & Conclusions

• **RSM proved to be an effective & efficient sampling method**
  – 25 points significantly less than 6400 exhaustive sampling
  – Even greater potential benefits in higher-dimensional space

• **Low order polynomial models were effective for the overall average and most of the sound classes, but not for all**

• **RSM approach seeks to develop simple, parsimonious models that:**
  – Adequately explore the hyperparameter space
  – Identify important factors and relationships
  – Note: RSM does not seek a global optimum
Summary & Conclusions

• RSM did a great job of estimating the results of the exhaustive sample

• RSM was able to achieve similar results to the other optimization methods, but with fewer samples and less computation time
  – RSM: 25 samples, ~45 minutes
  – Other optimization methods: 2000 samples, ~2.5 days
  – Exhaustive search: 6400 samples, ~8 days
Next Actions

- **Try RSM on larger hyperparameter problem**
  - More hyperparameters
  - More possible values
  - More complicated interaction between hyperparameters

- **Other flexible models**
  - Gaussian Process Models
  - Bayesian Treed Gaussian Process Models

- **Spectrograms performed mediocre and other time-frequency methods can/should be explored**
  - Mel-frequency cepstral coefficients
  - Wavelet scalograms
  - Hybrid/combined methods
Thank you for your time & attention!
Contact Information

Edward T. Nykaza, PhD
edward.t.nykaza@usace.army.mil

Program Manager RAPID Military Noise Environments
Project Leader, Research Engineer in Acoustics
ERDC Broadband Acoustics & Noise Group
US Army Corps of Engineers (USACE)
Engineer Research and Development Center (ERDC)
Construction Engineering Research Laboratory (CERL)
References

We would like to thank J. Salamon, C. Jacoby and J. P. Bello from New York University for compiling the dataset and making it freely available for non-commercial use. We would also like to thank FreeSound.org and its users for recording and hosting of wide variety of sound samples. For a complete attribution list of the dataset used in this study please visit https://serv.cusp.nyu.edu/projects/urbansounddataset/index.html.


Image source credit:
• http://i.huffpost.com/gen/3596508/images/o-SLEEP-NOISE-facebook.jpg
• http://www.flickr.com/photos/lingaraj/2415084235/sizes/l/
• https://upload.wikimedia.org/wikipedia/commons/thumb/3/3c/Qantas_b747_over_houses_arp.jpg/300px-Qantas_b747_over_houses_arp.jpg
• http://www.wikihow.com/images/5/5b/Choose-Headphones-Step-8.jpg
Backup Slides

The image shows a contour plot with two axes:
- The x-axis is labeled as x1 (nFFT), ranging from 1 to 8.
- The y-axis is labeled as x2 (nTimeBins), ranging from 1 to 8.

The plot contains color contours indicating different values, with the color bar on the right side showing values from 0.2 to 0.6. Notable contours include:
- A region in the top left corner with values around 0.40.
- A central region with a value of 0.25.
- A lower left region with values around 0.50.

The contour plot may represent a function or relationship between x1 and x2, with varying intensities indicating different levels of the function's output.
How does RSM compare to exhaustive sample?

Variable Importance (All Data)

Variable Importance
RSM

Variable Importance
(Calculated from Random Forest)
Best Values for Hyperparameters
Using Exhaustive Sample
Accuracy of y6 (Jackhammer)
Profiles at One Choice of Recommended Hyperparameters

![Graph showing profiles at one choice of recommended hyperparameters. The graph plots y4 (Drilling) against a range of x1 (nFFTs), x2 (nTimeBins), and x3 (% Overlap). The values are shown with error bars.](image)

**Link to real-time profiler**
y4 (Drilling)
Surface Plot at x3=50%
For these 2 Sound Classes and the Average, considering Computational Time, we might conclude:

- $X_1 = 8, X_2 = 1, X_3 = 10\%$ is a reasonable trade-off, however higher performance can be obtained for each Class.

Note: For the 4th order model, the classification accuracy predictions can appear nonsensical (greater than 1), however the shape of the response surface should be representative (see center tile, $y_6$ best, $x_2$ plot).